Enhancing Ad-hoc Relevance Weighting Using Probability Density Estimation

Xiaofeng Zhou
Information Retrieval and Knowledge Management Research Lab
Department of Mathematics and Statistics
York University
Toronto, Canada
bluesky@mathstat.yorku.ca

Xiangji Huang
Information Retrieval and Knowledge Management Research Lab
School of Information Technology
York University
Toronto, Canada
jhuang@yorku.ca

Ben He
Information Retrieval and Knowledge Management Research Lab
School of Information Technology
York University
Toronto, Canada
benhe@yorku.ca

ABSTRACT

Classical probabilistic information retrieval (IR) models, e.g. BM25, deal with document length based on a trade-off between the Verbosity hypothesis, which assumes the independence of a document's relevance of its length, and the Scope hypothesis, which assumes the opposite. Despite the effectiveness of the classical probabilistic models, the potential relationship between document length and relevance is not fully explored to improve retrieval performance. In this paper, we conduct an in-depth study of this relationship based on the Scope hypothesis that document length does have its impact on relevance. We study a list of probability density functions and examine which of the density functions fits the best to the actual distribution of the document length. Based on the studied probability density functions, we propose a length-based BM25 relevance weighting model, called BM25L, which incorporates document length as a substantial weighting factor. Extensive experiments conducted on standard TREC collections show that our proposed BM25L markedly outperforms the original BM25 model, even if the latter is optimized.

Categories and Subject Descriptors

H.4 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation, Theory

Keywords

Probabilistic IR, BM25, Document length, Normalization

1. INTRODUCTION

SIGIR'11 July 24-28, 2011, Beijing, China

An Information Retrieval system receives a query from user and returns the supposedly relevant documents. A crucial issue underlying an IR system is to rank the returned documents by decreasing order of relevance. Generally, ranking is based on a *weighting model*. The basic probabilistic model proposed in [23] is one of the most popular weighting models in modern IR systems, which is developed from the Probability Ranking Principle [22]. This probabilistic approach is refined based on the Verbosity hypothesis [24] which assumes independence of document length from relevance. In other words, long documents simply use more words than short documents to cover similar scope [24].

An opposite assumption about document length is the so-called Scope hypothesis, which states that some documents may contain more material than others if longer [24]. That is, long documents are more likely to be retrieved. In practice, a document may be considered as a trade-off between the Verbosity hypothesis and the Scope hypothesis. How to balance between these two hypotheses by modeling document length within the basic probabilistic weighting paradigm remains a challenging research issue. The impact of document length on relevance is particularly important for ad-hoc retrieval, where relevance is defined in a binary or graded manner. Compared to a short document, a long document is likely to be relevant if it contains paragraphs that meet the information need of the query, even if a large part of the document is in fact non-relevant.

To address the effect of document length on relevance, the basic probabilistic weighting function takes into account the document length d as follows [24]:

$$w(x,d) = \log \frac{P(\underline{x},d|R)}{P(\underline{x},d|\overline{R})} \frac{P(\underline{0},\Delta|\overline{R})}{P(\underline{0},\Delta|R)} \quad (1)$$

where $w(x,d)$ is the relevance weight of a given document. d is the document evidence for relevance, which is given by document length. Δ denotes the average document length of the reference vector $\underline{0}$, and $\underline{x}$ represents all other information about the document. $\overline{R}$ and R stand for the non-relevance and relevance, respectively. This function measures the difference between the probabilities of document length and all other information we have for the document when it is relevant and when it is not relevant, respectively, in log scale. The above equation also implies a relevant document should

receive a higher weight than a non-relevant document in order to achieve a satisfying retrieval performance.

Equation 1 can be further decomposed into the three components as follows [24]:

$$w(x,d) = w(\underline{x},d)_1 + w(\underline{0},d)_{21} + w(\underline{0},d)_{22} \quad (2)$$

where

$$w(\underline{x},d)_1 = \log \frac{P(\underline{x},d|R)}{P(\underline{x},d|\overline{R})} \frac{P(\underline{0},d|\overline{R})}{P(\underline{0},d|R)},$$

$$w(\underline{0},d)_{21} = \log \frac{P(\underline{0}|d,R)}{P(\underline{0}|d,\overline{R})} \frac{P(\underline{0}|\Delta,\overline{R})}{P(\underline{0}|\Delta,R)}$$

and

$$w(d,\Delta)_{22} = \log \frac{P(d|R)P(\Delta|\overline{R})}{P(d|\overline{R})P(\Delta|R)}.$$

Under the Verbosity hypothesis, document length has been considered as an independent evidence of relevance. This hypothesis nullifies the component $w(d,\Delta)_{22}$ in Equation 2, which as a consequence is set to zero [24]. Thus, the weighting function becomes

$$w(\underline{x},d) = w(\underline{x},d)_1 + w(\underline{0},d)_{21}. \quad (3)$$

The classical BM25 weighting model [10] is derived from Equation 3, more specifically, $w_{BM25} = w(\underline{x},d)$, where w_{BM25} is the relevance score of BM25, given by the following weighting function [10]:

$$w_{BM25} = \frac{(k_1+1)tf}{K+tf} * w^{(1)} * \frac{(k_3+1)qtf}{k_3+qtf} \oplus L \quad (4)$$

where

$$K = k_1 * \left((1-b) + b * \frac{dl}{avdl}\right),$$

$$w^{(1)} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)},$$

$$L = k_2 * nq * \frac{avdl-dl}{avdl+dl},$$

N is the number of indexed documents in the collection, n is the number of documents containing the query term, R is the number of known relevant documents to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf is within-query term frequency, dl is the document length (i.e. the document evidence d in Equation 2), $avdl$ is the average document length, nq is the number of query terms, k_i's and b are tuning constants (whose setting depends on the dataset used and is usually empirically determined), and $\oplus$ indicates that its following component is added only once per document. Particularly, b functions as a justification factor that adjusts the relative importance between the two hypotheses [25].

The main focus of this research is to study the relationship between document length and relevance in the context of the Scope hypothesis by exploring a list of probability density distribution for document length. The Scope hypothesis suggests the existence of a relationship between document length and relevance. It implies that the component $w(d,\Delta)_{22}$ in Equation 2 may not be zero. In this paper, we consider document length itself as a direct predictor of relevance. Our study of document length is based on the intuition that long documents tend to have high retrieval probabilities, since long documents usually have a large number of unique terms, which are likely to be picked up by the query term matching [28]. Our experiments show that long documents tend to have high frequencies of query terms, which leads to high relevance scores. This provides evidence supporting the Scope hypothesis that there exists a dependency of relevance on document length. Particularly, there are two extreme cases that need to be considered. One is that the relevance of each document in a collection is independent of its length. This is the case of BM11 or $b = 1$ in BM25 [25]. The other one is BM15 or $b = 0$ in BM25 [25], which implies that long documents are likely to be retrieved. To balance between these two extreme cases, Robertson et al. introduce a parameter b in BM25 to control the effect of tf normalization [25]. In our study, we first use statistical tools to learn the pattern of the relationship between the document length and its relevance, investigate the behavior of $w(d,\Delta)_{22}$ in Equation (2), and propose a new weighting function incorporating this relationship, which is the result of a mixture of the two hypotheses. Our results show that the retrieval performance of BM25 can be markedly improved over different settings of the parameter b by exploiting the document length evidence.

The rest of the paper is organized as follows. Section 2 gives a brief survey in previous work. Then Section 3 introduces the datasets and presents the idea of the length-based weighting function, and proposes seven models based on the density analysis. The proposed models are evaluated through extensive experiments in Section 4. Finally, we conclude on the work and suggest future research directions in Section 5.

2. RELATED WORK

The classical probabilistic models for IR rank documents according to their relevance scores, assigned by matching the query terms with adjustment for the relationship between document length and term frequency. This approach is developed based on the Verbosity hypothesis which assumes the document's relevance is independent of its length [24]. However, in practice, the impact of document length on relevance may be a mixture of both the Scope hypothesis and the Verbosity hypothesis [24].

Many previous studies have been conducted to investigate the impact of document length on relevance. Singhal *et al.* [28] suggest that long documents tend to have more unique terms, and consequently, long documents have a better chance to be retrieved than short documents. As the document length increases, the number of times the query terms occur in the documents also increases, which in turn increases the matching score. For instance, Singhal *et al.* illustrate that the probability of a document's relevance increases proportionally with document length in the early TREC test collections [28, 29]. Similar results have also been reported on the later "ad-hoc" test collections [15]. Moreover, a number of empirical studies have provided statistical evidence supporting that the probability of a document's relevance to an information need is considered to be correlated with the length of the document. Kraaij *et al.* show that the probability of relevance is positively correlated with document length on a number of TREC ad-hoc and Web collections [16]. Singhal *et al.* state that the documents retrieved by a model produce a retrieval pattern by the distribution of the document length [28]. Huang *et al.* use functional curve to approximate the distribution of document length on the TREC data sets and conclude that the retrieval system can be improved through an appropriate document length function [12, 13]. Furthermore, proper term weighting strategies based on document length can also improve retrieval

performance [31]. For example, normalization techniques have been applied for each term in the query through the length adjustment to avoid the bias introduced by document length. Normalizing the document length within a retrieval system could improve the performance [7]. By applying statistical regression of the similarity scores within the normalizing document length and query size, Lamprier *et al.* show that a significant improvement can be made to IR systems [17]. Blanco *et al.* device a probabilistic document length prior for language modeling [4]. Losada *et al.* apply smoothing techniques for document length in language modeling to show the significant impact of document length on the information retrieval performance [19]. They also argue that the relationship of document length and its relevance may not exist when the test collections are incomplete, although the evidence is not concrete enough to nullify the effect of document length on relevance [20]. For the issue addressed in [20], in this paper, we set up the experiments and use Expectation-maximization (EM) algorithm and bootstrapping method to avoid this problem [8].

3. DENSITY ANALYSIS AND LENGTH RELEVANCE WEIGHTING

Under the Scope hypothesis, $w(d, \Delta)_{22}$ in Equation 2 is no longer zero since a dependence of relevance on document length is assumed. To add the length information into the weighting function $w(x, d)$, we decompose the $w(d, \Delta)_{22}$ further into

$$w(d, \Delta)_{22} = \log \frac{P(d|R)}{P(d|\overline{R})} + \log \frac{P(\Delta|\overline{R})}{P(\Delta|R)} \quad (5)$$

The second component of Equation 5 is constant over a given document collection. This is because the average document length Δ for the reference vector $\underline{0}$ in a document collection is known and fixed. Therefore, for each document in a collection, the second component of Equation 5 above is the same across the whole document collection and does not affect the document ranking. For simplicity, we refer $w(d, \Delta)_{22}$ to as the first component in the Equation 5. Thus, the relevance weight $w(d, \Delta)_{22}$ is given by the log-odd of the relevance and non-relevance probabilities $P(d|R)$ and $P(d|\overline{R})$. In other words, $w(d, \Delta)_{22}$ measures the difference between the probabilities of given document length condition on relevance and non-relevance in the log scale. We name Equation 5 as *length relevance weighting*. Our ultimate goal is to calculate the $w(d, \Delta)_{22}$ in Equation 5, this needs a way to estimate the probabilities $P(d|R)$ and $P(d|\overline{R})$. By adding the measurement of document length itself into the basic weighting function, the retrieval system is expected to achieve high accuracy since the length information brings more evidence of relevance. The estimation of probability distribution function [1] of document length will be discussed in the next subsections.

The density estimation has three parts. First, we study the distributional pattern of document length on standard TREC test collections using kernel density estimation method [3, 27], which gives us a guidance in density estimation. In the second part, we apply data transformation techniques on the document length in order to get a better fitting of the document length distribution. Finally, Maximum Likelihood Estimation (MLE) is applied to obtain the distribution parameters estimates of document length. Length relevance weighting function is then derived based on the above density estimation.

In the rest of this paper, we use **d** to denote the document length. As a general rule, we usually make an assumption about observed d's, *i.e.* $d_1, d_2, ..., d_N$ are independent and identically distributed, N is the number of documents in the collection. We first introduce the test collections we used and then give the details of density estimation in the following subsections.

3.1 The TREC Test Collections

We examine the impact of document length on relevance using 4 standard TREC test collections. These four test collections are the most recent TREC datasets, and provide a good coverage on the a variety of commonly used datasets in IR evaluation, and are used for different test purposes and vary in size in term of the document length. Basic information about the test collections and topics are given in Table 1.

Table 1: Information about the test collections.

Coll.	TREC Task	Topics	# Docs
disk1&2	1-3, Ad-hoc	51-200	741,856
WT10G	9, 10 Web	451-550	1,692,096
.GOV2	2004-2006 Terabyte Ad-hoc	701-850	25,178,548
ClueWeb B	2009 Relevance Feedback	rf.01-rf.50	49,375,681

The disk1&2 collection contains newswire articles from various sources, such as Association Press (AP), Wall Street Journal (WSJ), Financial Times (FT), etc., which are usually considered as high-quality text data with little noise. It usually used for ad hoc test. The WT10G collection is a medium size crawl of Web documents, which was used in the TREC 9 and 10 Web tracks. It contains 10 Gigabytes of uncompressed data. The .GOV2 collection, which has 426 Gigabytes of uncompressed data, is a crawl from the .gov domain. This collection has been employed in the TREC 14 (2004), 15 (2005) and 16 (2006) Terabyte tracks. The ClueWeb collection is a very large crawl of the Web, and is currently the largest TREC test collection. We use the category B of ClueWeb, which contains about 50 million English Web pages, and its associated topics used in the TREC 2009 Relevance Feedback track. We index all documents in the above four collections. For all four test collections used, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed.

3.2 Kernel Density Analysis

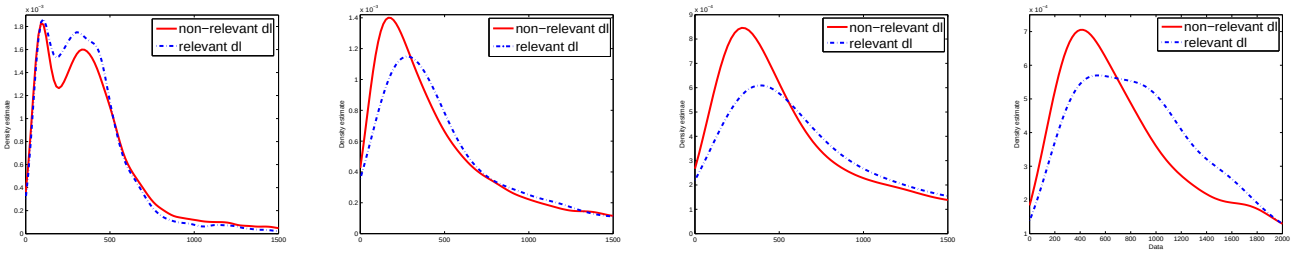
Kernel density estimation (or Parzen window method) is a non-parametric way of estimating the probability density function of a random variable. It can be used to extrapolate the data to the entire population as follows [27]:

$$\widehat{f_h}(d) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{d - d_i}{h}) \quad (6)$$

where $d_i, i = 1, \ldots, n$ is the independent and identically-distributed sample from some unknown distribution, n is the number of samples we draw from the population, K is the kernel function and h is the bandwidth (also called smoothing parameter). We can obtain the smoothing curve by adjusting the parameter h. Usually K is instantiated by a standard Gaussian function with a mean of zero and a variance of 1:

$$K(\frac{d - d_i}{h}) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(d - d_i)^2}{2h^2}) \quad (7)$$

[1]We use probability distribution function and probability density function interchangeable in the rest of this paper.



(a) disk1&2 (b) WT10G (c) .GOV2 (d) ClueWeb B

Figure 1: Kernel density estimate constructed from the document length on four test collections. "dl" stands for document length

Table 2: Basic statistical Information about the document length of four test collections. "rel" stands for relevant document length, "non-rel" stands for non-relevant document length.

Dataset	Mean		Standard Deviation		Minimum		Maximum	
	rel	non-rel	rel	non-rel	rel	non-rel	rel	non-rel
disk1&2	534.22	716.16	3960.11	3498.03	4	3	252983	252983
WT10G	1213.29	1165.83	2392.08	3801.12	24	5	48363	214693
.GOV2	3153.02	2821.13	4844.15	5022.19	11	10	47815	59566
ClueWeb B	1563.18	1583.32	1806.17	1888.64	50	52	16088	16087

Kernel density estimation gives us a global picture of the given dataset.

Figure 1 shows the distributional pattern of relevant and non-relevant document length for the test collections disk1&2, .GOV2, WT10G and ClueWeb B respectively using kernel density estimation with the standard normal kernel function. In Figure 1, the length curves have been cut when the document length is large than 1500 in $x-$axis in order to visualize the difference between the relevant and non-relevant document length because the long documents have low probabilities to be retrieved. The length of non-relevant and relevant documents from four test collections are both positively skewed and have a long tail with different tail shapes. Non-relevant document has higher frequency than relevant documents when the document length is relatively short. The curve on disk1&2 appears to have two distinct peaks called bimodality. In contrast, there is only one mode that arises on WT10G, .GOV2 and ClueWeb B.

We also list the basic statistics in Tables 2 for the length of relevant and non-relevant documents on each of the four test collections. At first glance, we may not see much difference between relevant and non-relevant documents from these two tables. However, further looking into these two tables, we observe that disk1&2 has the maximum relevant document length among the four test collections, the relevant and non-relevant document length of WT10G differ most among the four test collections, and ClueWeb B has the minimum length of both relevant and non-relevant documents.

We apply the data transformation techniques to visualize the difference between the relevant and non-relevant document length on each test collection used. By applying the data transform technique, we can also obtain higher likelihood distribution function and achieve more accurate estimates of distribution parameters.

3.3 Data Transformation

The data transformation technique is widely used in data processing or pre-processing for stabilizing the variance and make the data more normal distribution-like. In our case, all document lengths are positive, whose distribution is skewed to the right as described in Figure 1, and document length cannot be described by standard statistical methods because of the skewness. Therefore, data transformation is required to extract a better characteristic of the data. We initiate it by applying central limit theory, *i.e.* transformation I, it is also called standardization. By standardizing the data, it forces the data to locate on the common scales to be compared. Secondly, the Power transformation is from the family of functions that are applied to create a rank-preserving transformation of data which improves the correlation between variables and for other data stabilization procedures [5]. Box-Cox power transformation is commonly used to alleviate heteroscedasticity when the distribution of the variable of interest is not known, *i.e* transformations II and III. Transformation II is the special case of transformation III when $\theta = 0$. We also transform the document length to be within the scale of $0-1$ using transformation IV. The four types of transformation is described as follows:

- Transformation I: Standardization
$$z = \frac{d - \bar{d}}{s_d} \tag{8}$$
- Transformation II: Log transformation
$$z = \log(d) \tag{9}$$
- Transformation III: Box-Cox transformation
$$z = \frac{d^\theta - 1}{\theta} \tag{10}$$
Box-cox transformation is a parametric power transformation technique in order to reduce anomalies such as non-additivity, non-normality and heteroscedasticity, $\theta \neq 0$ is the transformation power.
- Transformation IV: Normalization
$$z = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{11}$$

where z is the document length after the transformation, $\bar{d}$ is the average document length, s_d is the standard deviation of document length, d_{min} denotes the minimum document length and d_{max} is the maximum document length.

Figure 2 plots the distributional pattern of transformed relevant and non-relevant document length on four test collections. In Figure 2, we examine length distribution patterns of relevant and non-relevant documents on the four test collections used. A major observation is that the the curves of the transformed document length distribution have similar shapes before and after the transformation. That is, the curves on disk1&2 remain bimodal, while the curves on the other three test collections are still left-skewed, but not as much as those of the original document length distribution, thanks to the data transformation. Moreover, the curves on .GOV2 and ClueWeb B become more symmetric after the transformation. On disk1&2, .GOV2 and WT10g, the center of the non-relevant document length distribution shifts far away to the right of the relevant document length distribution. From Figure 2(d), we can clearly see that relevant and non-relevant document length on Clueweb B can be distinguished from their distributional frequencies. Note that similar observations can also be drawn from other three collections used, but the difference between relevant and non-relevance document length distribution is not as obvious as on ClueWeb B. This is an encouraging finding as it gives us clue of differentiating between relevance and non-relevant documents based on their length distribution. In the next section, we propose to fit the document length distribution with a list of statistical distributions, in order to find the distributions that can match the characteristics of relevance and non-relevant documents.

3.4 Distribution of Document Length

The criterion of selecting distributions is that the distribution must be positive skewed with shape and rate parameters. With different shape and rate parameters, the probability distribution can describes as many different shapes as possible that document length may have. The commonly used distributions we applied to fit the transformed document length are as follows:

- Gamma distribution with $(\gamma > 0, \beta > 0)$
$$f(z) = \frac{z^{\gamma-1}e^{-z/\beta}}{\beta^{\gamma}\Gamma(\gamma)} \tag{12}$$
for $z \geq 0$, where γ and β are shape and scale parameters respectively. Varying setting of γ can lead to symmetrical or skewed figures.

- Normal distribution with (μ, σ), which is symmetric with respect to its mean value (μ). A Normal distribution is bell shaped and the shape is independent of its distribution parameters. The reason of choosing normal distribution is that transformation I try to standardize the document length. The Normal distribution density function is given as follows:
$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\{-\frac{(z-\mu)^2}{2\sigma^2}\} \tag{13}$$
where σ is the standard deviation.

- Lognormal distribution with (μ, σ)
$$f(z) = \frac{1}{z\sigma\sqrt{2\pi}}\, exp\{-\frac{(\log z-\mu)^2}{2\sigma^2}\} \tag{14}$$
for $z \geq 0$, where if Z is distributed lognormally with parameters μ and σ, $\log(Z)$ is distributed normally with a mean of μ and a standard deviation of σ. Lognormal and gamma distribution can produce similar graphs, but the curvature of lognormal distribution is more steep than gamma distribution.

- Inverse Gaussian distribution (IGD) with (μ, λ)
$$f(z) = \sqrt{\frac{\lambda}{2\pi z^3}}exp\{-\frac{\lambda}{2\mu^2 z}(z-\mu)^2\} \tag{15}$$
for $z > 0$, where $\mu > 0$ is the mean and $\lambda > 0$ is the shape parameter, changing λ changes the level of the skewness for the IGD.

- Weibull distribution with (a, b)
$$f(z) = \frac{b}{a}(\frac{z}{a})^{b-1}exp\{-(\frac{z}{a})^b\} \tag{16}$$
where $a > 0$ is the scale parameter and $b > 0$ is the shape parameter. Weibull distribution can produce the graph similar to Gamma distribution but with less steep curve.

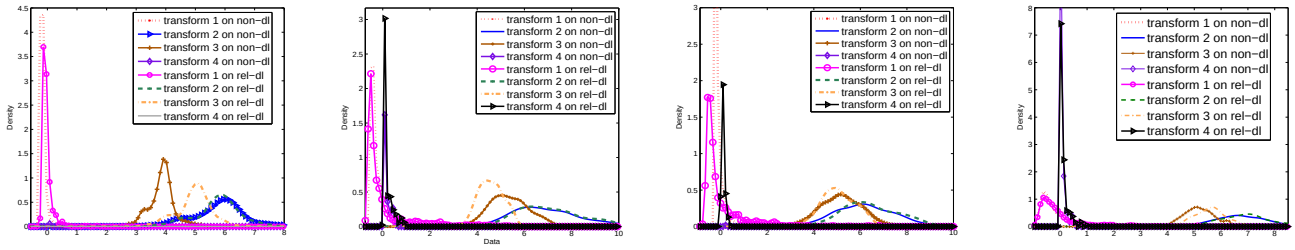
- Generalized Extreme Value distribution (GEV) (κ, μ, σ),
$$\begin{aligned} f(z) =& \frac{1}{\sigma} * exp\{-(1+\kappa\frac{(z-\mu)}{\sigma})^{-\frac{1}{\kappa}}\} \\ & * (1+\kappa\frac{(z-\mu)}{\sigma})^{(-1-\frac{1}{\kappa})} \end{aligned} \tag{17}$$
where $\kappa \neq 0$ is the shape parameter, μ is the location parameter and $\sigma > 0$ is the scale parameter. Compared to the statistical distributions mentioned above, GEV is a complicated distribution developed within the extreme value theory [9].

Figure 3 illustrates the six distribution fittings for the relevant document length of four test collections using Transformation I. Similar plots can be obtained for the relevant and non-relevant document length of all four test collections using Transformation I, Transformation III and Transformation IV respectively. All six distributions fit the .GOV2, WT10G and ClueWebB well, not disk1&2 since the bimodality. Inverse Gaussian ans GEV distribution fit the data best on all test collections, Weibull distribution can preserve the skewness better than the Lognormal, Gamma distribution, but normal distribution performs very badly in this case since skewness of the data. After the density functions are fit to the actual length distribution, it is necessary to use goodness of fit test to determine how well the distributions fit to the actual data.

3.5 Parameter Estimation

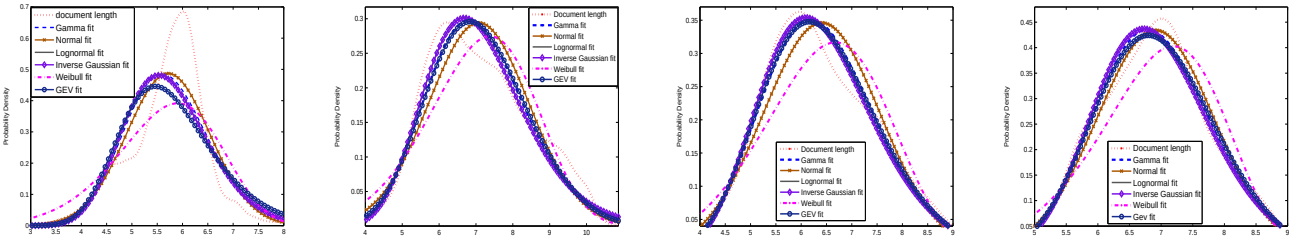
We adopt two methods in distribution parameter estimation for $P(d|R)$ and $P(d|\bar{R})$ in the Equation 2: bootstrapping and expectation-maximization(EM) algorithm. These are very simple but powerful statistical methods in parameter estimation.

Bootstrapping is a Monte Carlo method to learn about the sample characteristics to infer the population by resampling. It has been proved effective in reducing the bias of samples [8]. Adèr recommend to use bootstrapping when the sample size is insufficient for straightforward statistical inference [1]. The bootstrapping procedure is described as follows:



(a) disk1&2 (b) .GOV2 (c) WT10G (d) ClueWebB

Figure 2: Kernel density estimates constructed from the transformed document length of four test collections. "rel-dl" stands relevant document length and "non-dl" stands non-relevant document length.



(a) disk1&2 (b) .GOV2 (c) WT10G (d) ClueWebB

Figure 3: Distribution fittings for the relevant document length using transformation two $z = log(d)$.

1. Construct an empirical probability distribution Ω from the sample by placing a probability of $1/n$ at each point, $z_1, z_2, \ldots, z_n$ of the sample. This is the empirical distribution function of the sample, which is the nonparametric maximum likelihood estimate of the population distribution, ω. Now, each sample's element has the same probability of being drawn.

2. From the empirical distribution function, Ω, draw a random sample of size n with replacement. This step is called resampling.

3. Calculate the statistic of interest, θ, for this resample, yielding $\hat{\theta}^*$.

4. Repeat steps 2 and 3 for B times, where B is a large number, in order to create B resamples. The setting of B depends on the tests to be run on the data.

5. Compute $\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*$.

The Expectation-Maximization (EM) algorithm is a general algorithm for maximum-likelihood estimation where the data are "incomplete". The EM algorithm is an iterative method which seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

- Expectation step: Calculate the expected value of the log likelihood function with respect to the conditional distribution of y given z under the current estimate of the parameters $\theta^{(t)}$

$$Q(\theta|\theta^{(t)}) = E_{Y|z,\theta^{(t)}}[\log L(\theta; z, Y)]$$

- Find the parameter which maximizes following:

$$\theta^{(t+1)} = arg\max Q(\theta|\theta^{(t)})$$

where $L(\theta; z)$ is the likelihood function, θ is the parameter vector, z is the transformed document length and y represents the unobserved data.

The estimates from two methods are very close. For simplicity reason, we average the estimates from two methods in the experiments.

3.6 Length Relevance Weighting

After we obtain the distribution function for transformed document length Z, we apply the change variable technique [26] to obtain the distribution function for the original document length D, *i.e.* the document length before transformation. The theorem of change variable we used is as follows

THEOREM 1. *Let Z be a random variable with probability density function (pdf) $f_Z(z)$ and support $\mathcal{S}_Z$. Let $D = g(Z)$, where $g(z)$ is one to one differentiable function, on the support of Z, $\mathcal{S}_Z$. Denote the inverse of g by $z = g^{-1}(d)$ and let $\frac{\partial z}{\partial d} = \frac{\partial[g^{-1}(d)]}{\partial d}$. Then the pdf of D is given by*

$$f_D(d) = f_Z(z^{-1}(d))\,|\frac{\partial z}{\partial d}|, \quad for\ d \in \mathcal{S}_D \tag{18}$$

with the support of D which is the set $\mathcal{S}_D = \{d = g(z) : z \in \mathcal{S}_Z\}$.

Where $z^{-1}(d)$ is equivalent to Equations 8, 9, 10 and 11 when transforming the document length, and $|\frac{\partial z}{\partial d}|$ is the determinant of Jacobian of the transformation [11].

Based on the discussion above, we initiate the pattern of the document length by kernel density estimation. Based on the findings in step one, second, we apply data transformation and the change variable techniques to find the distribution functions of relevant and non-relevant document length and use MLE to obtain the parameter estimators. Two statistical methods, EM and bootstrapping are exploited to

prevent potential bias during parameter estimation, such as incomplete test collection, randomness of sampling. Hypothesis test employees to eliminate the distributions at 95% significant level. Finally Equation 5 is used to construct the following seven models:

1. In this model, transformed relevant and non-relevant document length follow the Normal distribution using same transformation I:standardization in Equation 8. We call it 'Normal'. Together, Using Equation 5 and change variable technique in Equation 18, the length relevance weighting function is as follows

$$w(d,\Delta)_{22} \propto -\frac{1}{2\sigma_1^2}\left[\left(\frac{d-\bar{d}}{s_d}\right)-\mu_1\right]^2+\frac{1}{2\sigma_2^2}\left[\left(\frac{d-\bar{d}}{s_d}\right)-\mu_2\right]^2 \tag{19}$$

where subscript 1 indicates that it is the estimates of distribution for relevant document length, 2 is the estimates of non-relevant document length distribution.

2. In this model, transformed relevant and non-relevant document length follow the Gamma distribution using transformation II:log transformation in Equation 9. We call it 'Log-Gamma'.

$$w(d)_{22} \propto (\gamma_1-\gamma_2)*\log d-\frac{\log d}{\beta_1}+\frac{\log d}{\beta_2} \tag{20}$$

3. In this model, transformed relevant and non-relevant document length follow the IGD distribution using transformation II:log transformation in Equation 9. We call it 'Log-IGD'.

$$\begin{aligned} w(d)_{22} \propto \log\sqrt{\frac{\lambda_1}{2*\pi*(\log d)^3}}-\log\sqrt{\frac{\lambda_2}{2*\pi*(\log d)^3}} \\ -\frac{\lambda_1*(\log d-\mu_1)^2}{2*\mu_1^2*\log d}+\frac{\lambda_2*(\log d-\mu_2)^2}{2*\mu_2^2*\log d} \end{aligned} \tag{21}$$

4. In this model, transformed relevant and non-relevant document length follow the Inverse Gaussian distribution distribution using transformation III:Box-Cox transformation in Equation 10. We call it 'Box-Cox-IGD'.

$$\begin{aligned} w(d)_{22} \propto \log\sqrt{\frac{\lambda_1}{2*\pi*(\frac{d^\theta-1}{\theta})^3}}-\frac{\lambda_1*(\log(\frac{d^\theta-1}{\theta})-\mu_1)^2}{2*\mu_1^2*\log(\frac{d^\theta-1}{\theta})} \\ -\log\sqrt{\frac{\lambda_2}{2*\pi*(\frac{d^\theta-1}{\theta})^3}}+\frac{\lambda_2*(\log(\frac{d^\theta-1}{\theta})-\mu_2)^2}{2*\mu_2^2*\log(\frac{d^\theta-1}{\theta})} \end{aligned} \tag{22}$$

5. In this model, transformed relevant and non-relevant document length follow the GEV distribution using transformation III:Box-Cox transformation in Equation 10. We call it 'Box-Cox-GEV'.

$$\begin{aligned} w(d)_{22} \propto &(-1-\frac{1}{\kappa_1})\log(1+\kappa_1\frac{V}{\sigma_1})-(1+\kappa_1\frac{V}{\sigma_1})^{-\frac{1}{\kappa_1}} \\ &-(-1-\frac{1}{\kappa_2})\log(1+\kappa_2\frac{V}{\sigma_2})+(1+\kappa_2\frac{V}{\sigma_2})^{-\frac{1}{\kappa_2}} \end{aligned} \tag{23}$$

where $V=\frac{d^\theta-1}{\theta}$.

6. In this model, transformed relevant and non-relevant document length follow the Lognormal distribution using transformation IV:normalization in Equation 11. We call it 'Lognormal'.

$$w(d)_{22} \propto -\frac{1}{2\sigma_1^2}\left[\log L-\mu_1\right]^2+\frac{1}{2\sigma_2^2}\left[\log L-\mu_2\right]^2 \tag{24}$$

where $L=(d-d_{min})/(d_{max}-d_{min})$

7. In this model, transformed relevant and non-relevant document length follow the Weibull distribution using transformation IV in Equation 11:normalization. We call it 'Weibull'.

$$w(d,\Delta)_{22} \propto (b_1-b_2)\log L-(\frac{L}{a_1})^{b_1}+(\frac{L}{a_2})^{b_2} \tag{25}$$

where $L=(d-d_{min})/(d_{max}-d_{min})$.

By adding $w(d,\Delta)_{22}$ into the weighting function $w(x,d)$ or BM25, we propose a new length-based weighting function BM25L as follows:

$$w(\underline{x},d)=(1-\beta)w_{BM25}\oplus\beta*w(d,\Delta)_{22} \tag{26}$$

where w_{BM25} is the relevance score of BM25, $\oplus$ indicate that the term $w(d,\Delta)_{22}$ is added only once for each document, β is not only the interpolation factor which is empirically determined and highly depends on the dataset used, but also an adjust factor of the mixture of two hypotheses: Verbosity and Scope hypothesis. A document could be either extreme or of mixture of these two hypotheses as discussed in[24]. More over, the reason of adding β here is that we ignore the constant term in the calculation of $\log\frac{P(\Delta|\overline{R})}{P(\Delta|R)}$, we need to adjust the scale for the weights between w_{BM25} and $w(\underline{x},d)$, and the weights between two hypotheses because BM25L consider the situation when both Verbosity and Scope hypothesis are presented in the same document. For a given query, each of the w_{BM25} or $w(\underline{x},d)$ scores is normalized by the maximum w_{BM25} or $w(\underline{x},d)$ score. The parameter β is obtained by Simulated Annealing [14] over a set of training topics.

4. EVALUATION

We introduce our methodology for evaluating the BM25L model in Section 4.1, and present the related evaluation results in Sections 4.2 and 4.3.

4.1 Evaluation Methodology

We evaluate our proposed BM25L model over the 4 test collections used, namely disk1&2, WT10G, .GOV2, and ClueWeb B. Each topic contains three topic fields, namely title, description and narrative. We only use the title topic field that contains very few keywords related to the topic. The title-only queries are usually short which is a realistic snapshot of real user queries in practice.

On each collection, the associated topics are divided into the odd-numbered and even-numbered topics. Over those two topic subsets, our proposed model is evaluated by a 2-fold cross-validation. In each fold, one of the topic subsets is used for training, and the other subset is used for testing purposes. More specifically, the half of the training topics with lower topic numbers are used to train the length distribution estimation parameters, and the other half of the training topics are used to train the score combination parameter β in Equation 26. Finally, our proposed BM25L model is evaluated by its retrieval performance on average over the two subsets of test topics. We use the TREC official evaluation measures in our experiments, namely the

statMAP on ClueWeb B [30], and the Mean Average Precision (MAP) on the other three collections [31].

Our evaluation baseline is the classical BM25 model with different settings of its parameter b. By varying the b value, we investigate to which extent BM25L improves the retrieval performance. In particular, we compare the retrieval performance of BM25L to BM25 with $b = 0$, that is, BM25 without tf normalization, and BM25 with its parameter b optimized. All statistical tests are based on Wilcoxon Matched-pairs Signed-rank test.

4.2 Comparison with BM25

Tables 3 and 4 compare the retrieval performance of BM25L to the original BM25 without tf normalization (i.e. when $b = 0$), and with the tf normalization with its parameter b optimized, respectively.

From Tables 3 and 4, we see that modeling document length distribution using GEV distribution leads to the most stable retrieval performance of our proposed length-based BM25L model. This is not of a surprise as we have shown that the GEV density fits the best to the actual document length distribution. Using the GEV density fitting of the document length, BM25L appears to outperform the BM25 baseline, and the improvement is statistically significant in most cases on all four test collections except WT10G.

The use of other distribution functions, in particular Gamma distribution, also leads to retrieval performance over the BM25 baseline on some of the test collections. However, their retrieval performance does not appear to be as robust as that obtained by GEV distribution. An extreme case is Normal distribution, which does not improve the BM25 baseline on disk1&2, WT10G, and .GOV2 when the parameter b is optimized. In contrast, on ClueWeb B, it provides an MAP that is as high as 0.6340, which is 173% higher than the BM25 baseline, even if its parameter b is optimized. Similar observation can also be made with Gamma distribution, which leads to a 130% improvement. One possible explanation is that it has lightest skewness among other 3 test collections that could be observed in Figure 3. Another possible reason for BM25L's extremely high retrieval performance with Gamma and Normal distribution on ClueWeb B is the shallow pool depth of this collection. Out of the four test collections used, ClueWeb B has the most incomplete relevance assessments, for which only the top-10 documents returned by the TREC participating runs are judged by human assessors [6]. As the top ranked documents are mostly overlong, the biase towards long documents in the document ranking could be so evident that the length distribution of relevant and non-relevant documents fits very well with the distribution functions on both training and testing topics. As a consequence, BM25L leads to extremely high retrieval performance on ClueWeb B.

To visualize the improvement brought the proposed length-based BM25L model, we plot the results in Figures 4 and 5 for the comparison to BM25 with $b = 0$ and with b optimized, respectively. As we can see on the WT10G collection, although the improvement is not as much as that obtained on other three test collections using all six distributions, the increase in retrieval performance is the evidence of length effect in information retrieval. Using Weibull distribution and normalization transformation has the best results, this may due to that Weibull distribution does retain the skewness of

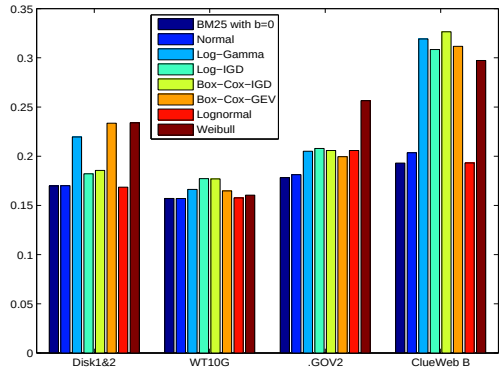


Figure 4: Performance of BM25L over BM25 with b=0

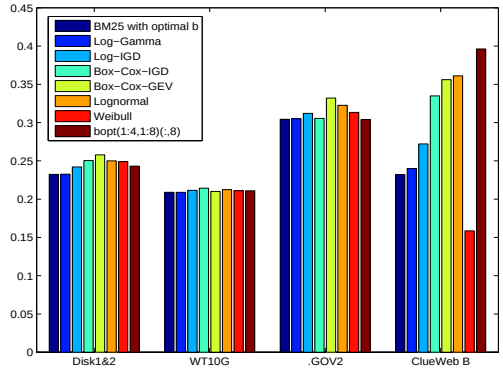


Figure 5: Performance of BM25L over BM25 with optimal b value

data between zero and one scale on all collections very well as we can see from Figure 3.

By comparing the performance improvement by BM25L over BM25 with $b = 0$ with the improvement over BM25 with the optimized b, we can see that the improvement over the optimized b is overall of a less scale than that over BM25 without tf normalization. This is because optimizing the parameter b in BM25 has exaggerated the length impact on the relevance weighting of term frequency tf, and in return, it reduces the impact of length relevance weighting itself on improving the document ranking.

When comparing BM25L with the best known results, for the WT10G, BM25L's best MAP is 0.2143, and the best published MAP is 0.2085. A possible explanation of the relatively minor improvement is as follows: the data transformation on WT10G does not show much difference between the length distribution of relevance and non-relevance documents. Compared to large-scale collections such as GOV2 and ClueWeb B, it leaves little room for the BM25L model to further improve the retrieval performance by utilizing such difference (in document length distribution). In other words, the TREC pools are biased by the length distribution. Such bias is minor on WT10G, and becomes evident on heterogeneous collections like GOV2 and ClueWeb B, which is captured by BM25L to boost the ranking effectiveness. For ClueWeb B, we believe the best statMAP in the TREC

Table 3: Evaluation results over the BM25 baseline with b=0. A star indicates a statistically significant improvement over the baseline.

Coll.	BM25	Normal	Log-Gamma	Log-IGD	Box-Cox-IGD	Box-Cox-GEV	Lognormal	Weibull
disk1&2	0.1698	0.1700	0.2195*	0.1821	0.1856*	0.2336	0.1685	0.2339*
WT10G	0.1571	0.1570	0.1663	0.1772*	0.1769	0.1647	0.1576	0.1604
.GOV2	0.1782	0.1812	0.2051*	0.2079*	0.2058*	0.1995	0.2058*	0.2564*
ClueWeb B	0.1930	0.2035	0.3192*	0.3084*	0.3265*	0.3117*	0.1931	0.2973*

Table 4: Evaluation results over the BM25 baseline with optimized setting of b. A star indicates a statistically significant improvement over this baseline.

Coll.	BM25	Normal	Log-Gamma	Log-IGD	Box-Cox-IGD	Box-Cox-GEV	Lognormal	Weibull
disk1&2	0.2324	0.2326	0.2421	0.2504*	0.2579*	0.2501*	0.2491*	0.2432*
WT10G	0.2090	0.2090	0.2115	0.2143	0.2101	0.2125	0.2111	0.2109
.GOV2	0.3044	0.3051	0.3121	0.3056	0.3321*	0.3227*	0.3134	0.3039
ClueWeb B	0.2322	0.2401	0.2722*	0.3350*	0.3561*	0.3612*	0.1586	0.3963*

2009 Relevance Feedback track, i.e. 0.2638, is achieved by combining BM25 with relevance feedback [32], although the overview paper is not available. Our model BM25L gives an MAP of 0.3963. For GOV2, on top of the retrieval baselines, e.g. BM25 and language model, the best run in TREC 2006 further improved the effectiveness by using pseudo relevance feedback and term dependency [18, 21]. Since our model only considers document length, the best MAP presented in this paper, i.e. 0.3321, is not directly comparable to the best known MAP of 0.3737. For disk1&2, there hasn't been known best result for all 150 topics used in the TREC 1-3 adhoc tasks. According to evaluatir.org, the best known MAP on each task is 0.2062, 0.2475 and 0.3231, respectively, with an average of 0.2589. Note that the above best known results are achieved by stacking additional techniques such as relevance feedback over the retrieval baseline. Therefore, the results are not directly comparable. Even though, BM25L provides an MAP of 0.2579.

4.3 Impact of Parameters

Experimental results in the previous section shows that, on one hand, BM25L leads to more improvement over BM25 when tf normalization is disabled. This is expected since there is no length information added to BM25 with $b = 0$ when compare to BM25L. On the other hand, BM25L provides higher MAP/statMAP values when the tf normalization parameter b is optimized. From this observation, a question arises: what is the impact of the setting of b on BM25L's effectiveness? To answer this question, Figure 6 plots the MAP/statMAP obtained by BM25L using the 6 different statistics of the document length distribution against different b values, from 0 to 1. BM25L and the original BM25's retrieval performance is seen to be correlated. A better setting of BM25's b leads to a better retrieval performance of BM25L. The document length itself do have a power as a stand-alone factor on the document relevance weighting other than normalization adjustment. The results for the full range of b are illustrated in Figure 6 for all 4 test collections.

Another important factor that could heavily affect BM25L's retrieval performance is the parameter β in Equation 26. Figure 7 plots the MAP/statMAP obtained by BM25L against β on the four collections used. As we can see that length impact on the relevance weighting increase first as β increase, then either decrease or remain flat as β increase. This is no coincidence because with only one factor, *i.e.* length, among other many important factors that can affect document relevance weighting, the improvement would be limited.

In summary, we have shown that the length information can be used for leveraging the bias towards long documents in the document ranking. The retrieval performance of the classical well-established BM25 model can be marked improved by incorporating a length-based weighting component with different settings of BM25's tf normalization parameter, including the optimal setting. Finally, we recommend applying GEV distribution for modeling the document length distribution as it has demonstrated effective and robust retrieval performance in our experiments. In our experiments, all parameters are learned from the training data in the two-fold cross-validation. It is therefore of note that our proposed model is trained and tested with different queries.

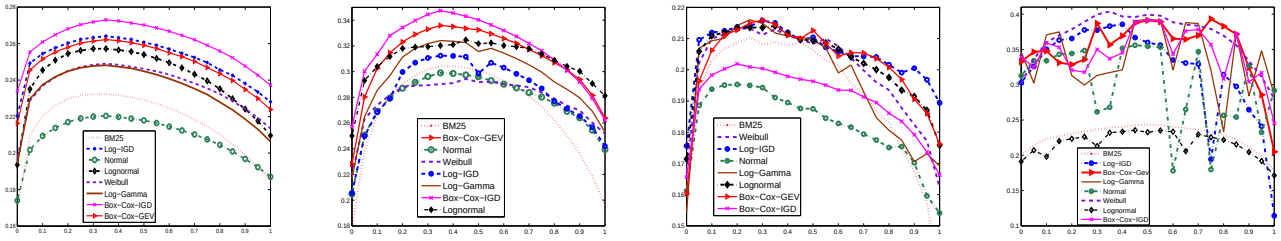
5. CONCLUSION AND FUTURE WORK

Our research in this paper is based on the assumption that a document may exhibit both Verbosity and Scope hypotheses. We derive the relationship between document length and its relevance through a list of probability density functions, and propose a BM25L model that incorporates this relationship into the classical BM25 model. The proposed BM25L model is evaluated on standard large-scale TREC collections. Our experiments demonstrate that BM25L is able to markedly outperform the BM25 baseline even with the optimized tf normalization. The results empirically confirm our assumption that the actual relationship between document length and relevance is a mixture and compromise between the Verbosity and Scope hypotheses.

In this paper, we have proposed a general method of utilizing the relationship between document length and relevance for improving retrieval performance. The proposed method is shown to be effective in its application to the classical BM25 weighting model. In the future, we plan to apply our proposed method to other state-of-the-art IR models, such as language modeling, or the PL2 model [2]. In addition, we also plan to investigate possible ways to balance between the effectiveness and efficiency of our proposed method.

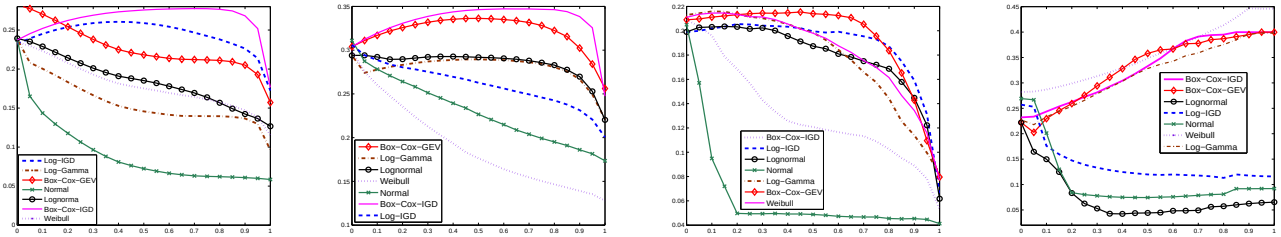
6. ACKNOWLEDGMENTS

This research is supported in part by NSERC of Canada and the Early Researcher Award/Premier's Research Excellence Award. We thank four anonymous reviewers for their excellent comments on this paper.



(a) disk1&2 (b) .GOV2 (c) WT10G (d) ClueWeb B

Figure 6: The MAP/statMAP values obtained against the parameter b



(a) disk1&2 (b) .GOV2 (c) WT10G (d) ClueWeb B

Figure 7: The MAP/statMAP values obtained against the parameter β

7. REFERENCES

[1] H. J. Adèr, G. J. Mellenbergh, and D. J. H. Huizen. *Advising on Research Methods: A Consultant's Companion*. The Netherlands: Johannes van Kessel.

[2] G. Amati and C. J. V. Rijsbergen. Probabilistic models for information retrieval based on measuring the divergence from randomness. *ACM Transaction on Information Systems(TOIS)*, 20(4):357–389, 2002.

[3] C. Bishop. *Neural networks for pattern recognition*. Oxford university Press, Oxford, UK, 1996.

[4] R. Blanco and A. Barreiro. Probabilistic document length priors for language models. In *Proceedinds of ECIR*, pages 394–405, 2008.

[5] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society.*, 26(2):211–252, 1964.

[6] C. Buckley and S. E. Robertson. Overview of trec 2009 relevance feedback track. In *Proceedings of TREC 2009*, 2009.

[7] A. Chowdhury, M. C. McCabe, D. Grossman, and O. Frieder. Document normalization revisited. In *Proceedings of the 25th ACM SIGIR Conference*, pages 381–382, 2002.

[8] J. E. Gentle. *Elements of Computational Statistics*. Springer, 2002.

[9] E. Gumbel. *Statistics of Extremes*. Columbia University Press, 1958.

[10] M. Hancock-Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. W. Williams. Okapi at TREC-5. In *Proceedinds of 5th Text Retrieval Conference*, pages –1–1, 1996.

[11] R. V. Hogg, hoseph W. Mckean, and A. T. Craig. *Introduction to mathematical statistics*. Pearson Education, Upper Saddle River, N.J., 2005.

[12] X. Huang, F. Peng, D. Schuurmans, N. Cercone, and S. E. Robertson. Applying machine learing to text segmentation for information retrival. *Information Retrieval*, 6(4):332–362, 2003.

[13] X. Huang, S. E. Robertson, N. Cercone, and A. An. Probability-based chinese text processing and retrieval. *Computational Intelligence: An International Journal (CI)*, 16(4):552–569, 2000.

[14] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[15] W. Kraaij and T. Westerveld. Tno/ut at trec-9: How different are web documents. In *Proceeding of 9th Text Retrieval Conference*, 2000.

[16] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of 25th ACM SIGIR conference*, pages 27–34, 2002.

[17] S. Lamprier, T. Amghar, B. Levrat, and F. Saubion. Document length normalization by statistical regression. In *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence*, pages 11–18, 2007.

[18] J. Li and H. Yan. Peking university at the trec 2006 terabyte track. *TREC2006*.

[19] D. E. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11(2):109–138, 2008.

[20] D. E. Losada, L. Azzopardi, and M. Baillie. Revisiting the relationship between document length and relevance. In *Proceedings of the 17th ACM conference*, pages 419–428, 2008.

[21] D. Metzler, T. Strohman, and W. Croft. Indri trec notebook 2006: Lessons learned from three terabyte tracks. *TREC2006*.

[22] S. E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33:294–304, 1977.

[23] S. E. Robertson, C. J. van Rijsbergen, and M. F.Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3rd ACM SIGIR conference*, pages 35–56, 1980.

[24] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th ACM SIGIR conference*, pages 232–241, 1994.

[25] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedinds of 3rd Text Retrieval Conference*, pages 109–126, 1994.

[26] J. Shao. *Mathematical Statistics*. Springer, 2003.

[27] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman&Hall, 1986.

[28] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of 19th ACM SIGIR Conference*, pages 21–29, 1996.

[29] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing and Management: an International Journal*, 32(5):619–633, 1996.

[30] E. M. Voorhees. Notebook of TREC 2009. 2009.

[31] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, Massachusetts, 2005.

[32] Z. Ye, X. Huang, B. He, and H. Lin. York university at the trec 2009: Relevance feedback track. *TREC2009*.